Viewing the Internet of Events through a Process Lens

Wil M.P. van der Aalst

Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands

Abstract. The spectacular growth of event data is rapidly changing the Business Process Management (BPM) discipline. It makes no sense to focus on process modeling (including model-based analysis and model-based process automation) without considering the torrents of factual data in and between today's organizations. Hence, there is a need to connect BPM technology to the "internet of events" and make it more evidence-based BPM. However, the volume (size of data), velocity (speed of change), variety (multiple heterogeneous data sources), and veracity (uncertainty) of event data complicate matters. Mainstream analytics approaches are unable to turn data in to insights, once things get more involved. Therefore, they tend to focus on isolated decision problems rather than providing a more holistic view on the behavior of actors within and outside the organization. Fortunately, recent developments in process mining make it possible to use process models as the "lens" to look at (low) level event data. Viewing the internet of events through a "process lens" helps to understand and solve compliance and performance related problems. In fact, we envision a new profession —the process scientist— connecting traditional model-driven BPM with data-centric approaches (data mining, statistics, and business intelligence). Process mining provides the process scientist with a powerful set of tools and prepares BPM for a highly connected world where processes are surrounded by devices emitting events.

1 Introduction

Organizations are competing on analytics and only organizations that intelligently use the vast amounts of data available will survive. Process-mining techniques enable the analysis of a wide variety of processes using event data. For example, event logs can be used to automatically learn a process model (e.g., a Petri net or BPMN model). Next to the automated discovery of the real underlying process, there are process-mining techniques to analyze bottlenecks, to uncover hidden inefficiencies, to check compliance, to explain deviations, to predict performance, and to guide users towards "better" processes. Dozens (if not hundreds) of process-mining techniques are available and their value has been proven in many case studies. See for example the twenty *case studies* on the webpage of the IEEE Task Force on Process Mining [7]. The growing number of commercial *process mining tools* (Disco, Perceptive Process Mining, Celonis

Process Mining, QPR ProcessAnalyzer, Software AG/ARIS PPM, Fujitsu Interstage Automated Process Discovery, etc.) further illustrates the uptake of process mining. The recent Massive Open Online Course (MOOC) on process mining attracted over 41.500 participants [4].

Process mining provides the interface between process models and event data. On the one hand, conventional Business Process Management (BPM) and Workflow Management (WfM) approaches and tools are mostly model-driven with little consideration for event data. On the other hand, Data Mining (DM), Business Intelligence (BI), and Machine Learning (ML) focus on data without considering end-to-end process models. Process mining aims to bridge the gap between BPM and WfM on the one hand and DM, BI, and ML on the other hand. Here, the challenge is to turn torrents of event data ("Big Data") into valuable insights related to process performance and compliance.

Fig. 1. Process models can be seen as the glasses through which one can see structure in otherwise puzzling event data.

This paper does not focus on specific process mining algorithms. Instead, it focuses on the interplay between event data and process models. As illustrated in Figure 1, process models can be used to view event data in such a way that actionable knowledge can be extracted. Process models can be used to extract real value from event data. However, this is only possible if model and data are aligned. This is where process mining plays a crucial role.

This paper coins the term "process lens" and demonstrates that processes can indeed be used to interpret confounding event data.

2 Internet of Events

In [3] the term *Internet of Events* (IoE) was coined to refer to all event data available. As described in [6, 8], society shifted from being predominantly "analog" to "digital" in just a few years. Society, organizations, and people are "Always On". Data is collected *about anything*, *at any time*, and *at any place*. *Event data* are the most important source of information. Events may take place inside a machine (e.g., an X-ray machine or baggage handling system), inside an enterprise information system (e.g., a order placed by a customer), inside a hospital (e.g., the analysis of a blood sample), inside a social network (e.g., exchanging e-mails or twitter messages), inside a transportation system (e.g., checking in, buying a ticket, or passing through a toll booth), etc. Events may be "life events", "machine events", or both.

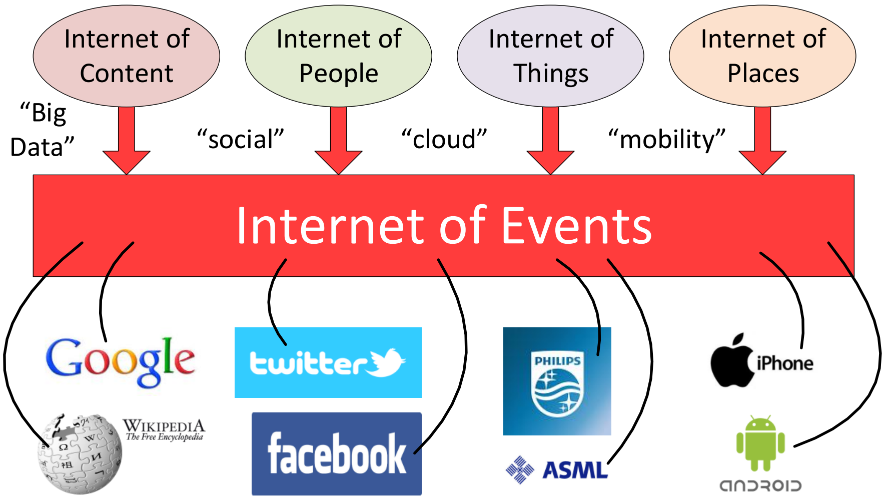


Fig. 2. The Internet of Events (IoE) is based on the Internet of Content (IoC), the Internet of People (IoP), the Internet of Things (IoT), and the Internet of Locations (IoL).

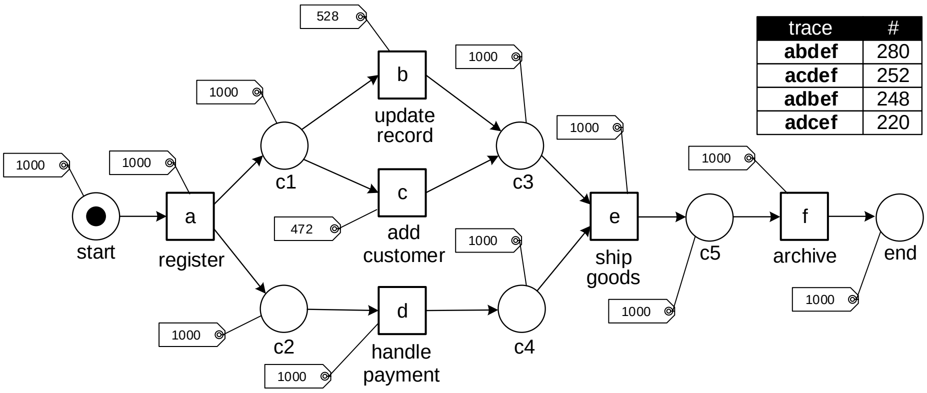
Figure 2 aims to characterize the types of events available for analysis. As shown the *Internet of Events* (IoE) is composed of:

- The *Internet of Content* (IoC): all information created by humans to increase knowledge on particular subjects. The IoC includes traditional web pages, articles, encyclopedia like Wikipedia, YouTube, e-books, newsfeeds, etc.
- The *Internet of People* (IoP): all data related to social interaction. The IoP includes e-mail, facebook, twitter, forums, LinkedIn, etc.
- The *Internet of Things* (IoT): all physical objects connected to the network. The IoT includes all things that have a unique id and a presence in an internet-like structure. Things may have an internet connection or tagged using Radio-Frequency Identification (RFID), Near Field Communication (NFC), etc.
- The *Internet of Locations* (IoL): refers to all data that have a spatial dimension. With the uptake of mobile devices (e.g., smartphones) more and more events have geospatial attributes.

Obviously, the IoC, the IoP, the IoT, and the IoL are partially overlapping. The spectacular growth of IoE impacts BPM, e.g., process improvements will increasingly be driven by analytics. At the same time, organizations have difficulties exploiting the data they have. Therefore, we propose *process models to be used as lenses to view data* that are otherwise confusing.

3 Process Lens

To use a process model as a "lens" to observe the data from a particular viewpoint, it is not sufficient to have a model and data. Both (i.e. model and data) need to be aligned. This can be achieved through process mining. Based on the event data, a model can be discovered that is automatically aligned with the data. Through conformance checking, it is possible to align normative models with the data and highlight discrepancies between modeled behavior and observed behavior.



trace	#
abdef	280
acdef	252
adbef	248
adcef	220

Fig. 3. A process model discovered for an event log with 1000 cases. The model is expressed in terms of a Petri net.

To introduce the notion of process discovery, we consider Figure 3, which shows an event log and a discovered process model. The event log holds information on 1000 cases: 280 cases followed the trace **abdef**, 252 followed the trace **acdef**, etc. The activity names have been shortened to a single letter, e.g., **a** = **register**. In fact, the event log in Figure 3 is significantly simplified by abstracting away many aspects. Normally, each *event* refers to a *case* and an *activity*. An event also has a *timestamp* and may refer to the *resources* used, there may be *transactional information*, and any number of attributes. In the example log such information is missing. Cases and events cannot be distinguished, but the compact representation allows us to illustrate the basics of process mining. There are 280 cases that followed the same sequence of activities: **abdef**. The event log in Figure 3 has 5000 events, e.g., 1000 **a** events that always happen first.

The discovered process model in Figure 3 is expressed in terms of a Petri net. In the initial state shown, only the transition having activity label **a** can occur. A transition (represented as a square) can occur if all input places (represented as circles) have a token (represented by a black dot). If a transition occurs, tokens are consumed from all input places and produced for all output places. After the occurrence of **a** in Figure 3, there are tokens in **c1** and **c2**. Hence after **a** either (1) **b** and **d** occur or (2) **c** and **d** occur (in any order). Then **e** occurs, followed by **f**. The end state is the state with a token in **end**. Note that the process model is able to replay all 1000 cases: they all start in the initial state and finish in the desired end state. Figure 3 also shows the number of times each place and transition is visited, e.g., **b** occurs 528 times.

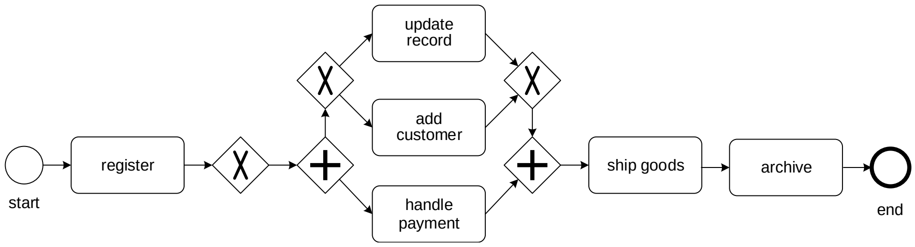
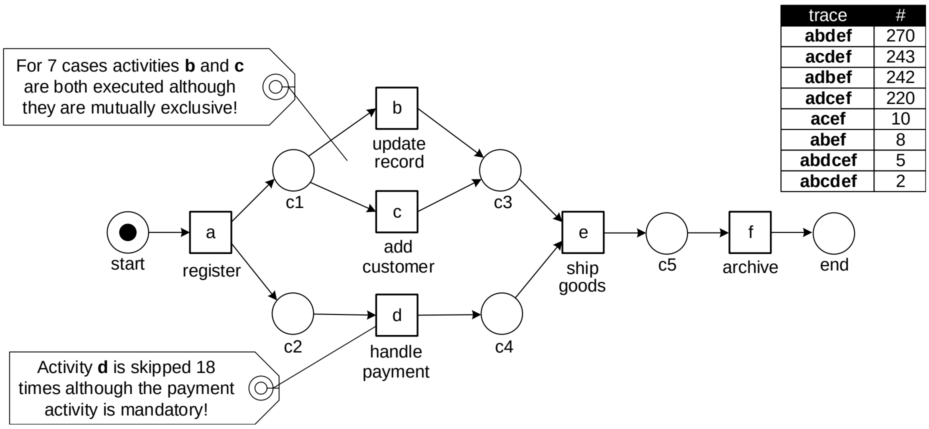


Fig. 4. A BPMN model corresponding to the discovered Petri net in Figure 3.

The BPMN model in Figure 4 has the same behavior as the Petri net model in Figure 3. We would like to stress that the notation is less relevant here: Automatic translations are possible and "observed behavior does not have a preference for a particular syntax" (although people like to believe differently). Only things that can be related to event data matter!

Process models can be discovered automatically or made by hand. In both scenarios, it is possible to *check conformance*. This is illustrated in Figure 5. Assume that sometimes **d** is skipped or both **b** and **c** occur. Hence, reality as described in the event log deviates from the model. Figure 5 shows some diagnostics. Conformance checking techniques will immediately reveal such deviations.



For 7 cases activities **b** and **c** are both executed although they are mutually exclusive!

Activity **d** is skipped 18 times although the payment activity is mandatory!

trace	#
abdef	270
acdef	243
adbef	242
adcef	220
acef	10
abef	8
abdcef	5
abcdef	2

Fig. 5. The event log now has 25 cases that do not fit into the normative process model. By replaying the event log one can see that activity **d** is sometimes skipped and activities **b** and **c** are both executed although the model does not allow for this.

Note that the process model is used as the "lens" to show non-conforming behavior.

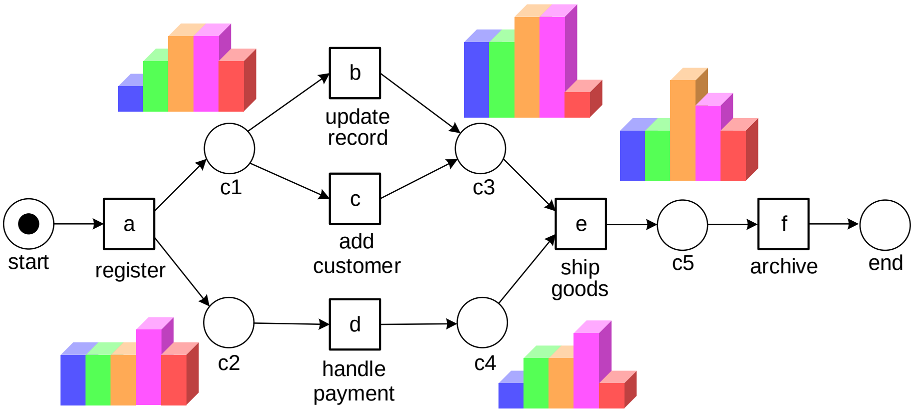


Fig. 6. By replaying the event log on a discovered process model, one can see where the bottlenecks are in the process.

Replay can also be used to show bottlenecks, see Figure 6. Note that the process model is now used as the lens to show performance-related behavior.

It is also possible to replay streaming event data, i.e., align cases to the model while they are still running. This can be used to show the "traffic" in the process, as illustrated by Figure 7. There are three types of cases (represented

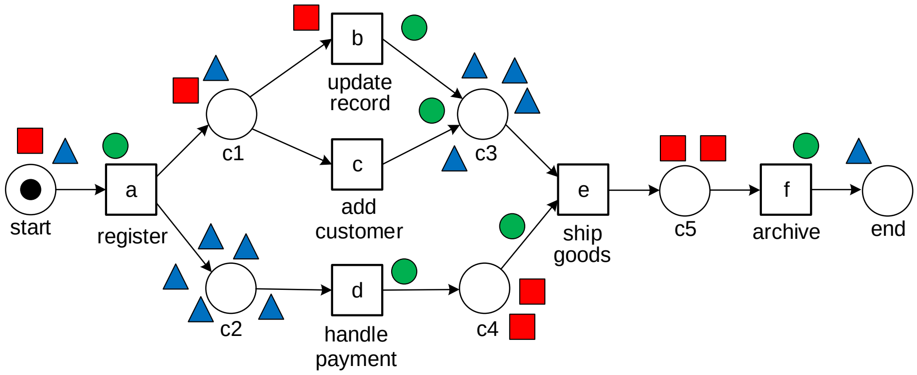
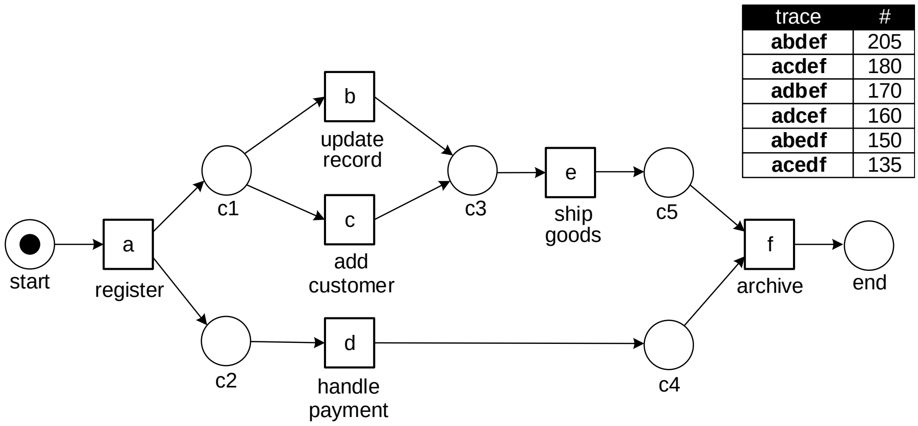


Fig. 7. By mapping running cases onto the model, one can see the “traffic jams” in an organization.

using triangles, squares, and circles). In Figure 7 one can see the congestion of particular case types at any point in time.



trace	#
abdef	205
acdef	180
adbef	170
adcef	160
abedf	150
acedf	135

Fig. 8. Due to concept drift, the model is changing. Hence, the model needs to be adapted continuously to avoid misleading diagnostics.

The event log in Figure 8 describes the same process, but in a later period. Again 1000 cases are recorded, but now the shipment (**e**) often occurs before the payment (**d**). This was not possible according to the original model. The Petri net shown in Figure 8 shows the updated process model also allowing for this new behavior.

The change from the process in Figure 3 to the process in Figure 8, illustrates that the simple dashboards and reports provided by contemporary Business Intelligence (BI) tools are inadequate. One needs to look *into* the process and cannot reduce reality to a few Key Performance Indicators (KPIs). Suppose one is interested in the delay between payments (activity **d**) and shipments (activity **e**). This corresponds to the sojourn time of tokens in place **c4** in Figure 3. The process could have been instrumented to measure these delays. However, after the drift these times can be negative. Shipments (activity **e**) happen before payments (activity **d**), and statistics can become very misleading. If activities are skipped, similar problems occur. Hence, it is vital to align event data with an up-to-date process model. Existing BI tools not supporting process mining, cannot cope with such issues. To use a BI tool, an idealized process is assumed and only high-level measurements are performed.

Process mining provides a way to look into the process and view event data in a process-centric manner. Process models provide the lenses to make sense of event data. The volume (size of data), velocity (speed of change), variety (multiple heterogeneous data sources), and veracity (uncertainty) of event data necessitates state-of-the-art techniques that are able to reliably and efficiently interpret recorded events.

4 Process Mining in the Large

The reader is referred to [1] for an introduction to process mining. Process mining extends far beyond process discovery. The alignment of process models and event data enables all kinds of analytics, e.g., decision point analysis, bottleneck analysis, time prediction, resource recommendation, and compliance checking.

The spectacular growth of event data provides numerous opportunities for process mining in any business. However, there are also challenges related to "process mining in the large". Fortunately, recent developments show that process mining is quite scalable compared to classical data mining techniques. Some examples:

- Many process-mining techniques (but obviously not all) are linear in the size of the event log. If the number of activities is limited, then the time need to discover a process model corresponds to the time to traverse the data.
- There are some discovery approaches that are more time consuming and computing alignments (e.g., for conformance checking or performance analysis) is known to be time consuming. Fortunately, there are generic techniques [2] to decompose large process mining into many smaller ones that can be solved much faster.
- There are many ways to distribute process-mining problems. Next to the process-mining specific approach in [2], many subproblems can be trivially distributed using MapReduce approaches exploiting for example Hadoop for distributed storage and distributed processing.
- Events logs can be decomposed for performance and scalability reasons. However, it is often also useful to partition the log and then compare the results. Techniques and tools for comparative process mining are emerging and essential for conducting process mining at a larger scale, e.g., for comparing different departments, regions, customer groups, or periods.

- The notion of "process cubes" can be used for comparative process mining. Events are stored in the cells of a multidimensional database. This is closely related to Online Analytical Processing (OLAP) technologies that aim to answer multi-dimensional analytical queries using operators such as slice, dice, roll-up, and drill-down.

The above developments illustrate that process mining fits well with other developments in the context of Big Data.

5 Towards a Process Scientist

Hal Varian, the chief economist at Google said in 2009: "The sexy job in the next 10 years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?". The later article with the provocative title "Data Scientist: The Sexiest Job of the 21st Century" [5] generated lots of attention for this new profession. Indeed, just like computer science emerged from mathematics in the 70-ties and 80-ties, data science is now emerging from computer science, statistics, and management science. However, let's not forget about processes. As argued in this paper, processes provide the lenses to look at event data from different angles. The focus on data analysis is good, but should not frustrate process-orientation. In the end, good processes are more important than information systems and data analysis. The old phrase "It's the process stupid" is still valid. Hence, we advocate the need for *process scientists* that will drive process innovations while exploiting the Internet of Events (IoE).

References

1. W.M.P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes.* Springer-Verlag, Berlin, 2011.
2. W.M.P. van der Aalst. Decomposing Petri Nets for Process Mining: A Generic Approach. *Distributed and Parallel Databases*, 31(4):471–507, 2013.
3. W.M.P. van der Aalst. Data Scientist: The Engineer of the Future. In K. Mertins, F. Benaben, R. Poler, and J. Bourrieres, editors, *Proceedings of the I-ESA Conference*, volume 7 of *Enterprise Interoperability*, pages 13–28. Springer-Verlag, Berlin, 2014.
4. W.M.P. van der Aalst. Process Mining: Data science in Action. Coursera Course, November 2014. https://www.coursera.org/course/procmin.
5. T.H. Davenport and D.J. Patil. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, pages 70–76, October 2012.
6. M. Hilbert and P. Lopez. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025):60–65, 2011.
7. IEEE Task Force on Process Mining. Process Mining Case Studies. `http://www.win.tue.nl/ieeetfpm/doku.php?id=shared:process_mining_case_studies`, 2013.

8. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, 2011.